NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

DETECTING POTENTIALLY COMPROMISED CREDENTIALS IN A LARGE-SCALE PRODUCTION SINGLE-SIGNON SYSTEM

by

Timothy Riley

June 2014

Thesis Advisor:	Robert Beverly
Second Reader:	John D. Fulp

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704–0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	06-20-2014	Master's Thesis

4. TITLE AND SUBTITLE

DETECTING POTENTIALLY COMPROMISED CREDENTIALS IN A LARGE-SCALE PRODUCTION SINGLE-SIGNON SYSTEM

5. FUNDING NUMBERS

6. AUTHOR(S)

Timothy Riley

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Postgraduate School
Monterey, CA 93943

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

The views expressed in this document are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol Number: N/A.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(maximum 200 words)*

We posit that potentially compromised credentials are detectable by analyzing the system artifacts of a large-scale production, single-signon system. With permission from the Defense Manpower Data Center, we analyze a year's worth of system artifacts produced by the Department of Defense Self-Service Logon system. Using industry standard tools and descriptive statistics we develop a repeatable process that identifies potentially compromised credentials. We look for characteristics that coincide with compromised credentials and evaluate our approach by obtaining the ground truth on several of the credentials we identify.

14. SUBJECT TERM

credential, compromised, single-signon, dslogon, elasticsearch, kibana, bigdata

15. NUMBER OF PAGES 81

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

DETECTING POTENTIALLY COMPROMISED CREDENTIALS IN A LARGE-SCALE PRODUCTION SINGLE-SIGNON SYSTEM

Timothy Riley
Civilian, Defense Manpower Data Center
BS, University of Mary Washington, 2001

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
June 2014

Author: Timothy Riley

Approved by: Robert Beverly
Thesis Advisor

John D. Fulp
Second Reader

Peter Denning
Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

We posit that potentially compromised credentials are detectable by analyzing the system artifacts of a large-scale production, single-signon system. With permission from the Defense Manpower Data Center, we analyze a year's worth of system artifacts produced by the Department of Defense Self-Service Logon system. Using industry standard tools and descriptive statistics we develop a repeatable process that identifies potentially compromised credentials. We look for characteristics that coincide with compromised credentials and evaluate our approach by obtaining the ground truth on several of the credentials we identify.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
API	Application Programming Interface
CAC	Common Access Card
CBL	Composite Blocking List
CDF	Cumulative Distribution Function
DEERS	Defense Enrollment and Eligibility Reporting System
DMDC	Defense Manpower Data Center
DNS	Domain Name System
DNSBL	Domain Name System (DNS) Block List
DOD	Department of Defense
DSL	DOD Self Service Logon
DVA	Department of Veterans Affairs
FIPS	Federal Information Processing Standards
HIV	Human Immunodeficiency Virus
HR	human resources
HSPD	Homeland Security Presidential Directive
IP	Internet Protocol
ISP	Internet Service Provider
IT	information technology
JSON	JavaScript Object Notation

NIST	National Institute of Standards and Technology
NoSQL	No Structured Query Language
PBL	Policy Block List
PII	personally identifiable information
PIP	Personnel Identity Protection
PTSD	Post-Traumatic Stress Disorder
RAPIDS	Real-time Automated Personnel Identification System
RDBMS	Relational Database Management System
REST	Representational State Transfer
RFC	Request For Comment
SMTP	Simple Mail Transfer Protocol
SOP	standard operating procedure
SP	special publication
SQL	Structured Query Language
SSH	Secure Shell
UBE	Unsolicited Bulk Email
U.S.	United States
USPS	United States Postal Service
XML	eXtensible Markup Language

Executive Summary

The Department of Defense (DOD) Self Service Logon (DSL) is a large-scale single-signon system, that issues username and password based credentials that adhere to the standards established in the National Institute of Standards and Technology *Electronic Authentication Guidelines Special Publication* 800-63-2 [1]. The DSL was established in 2008 by Defense Manpower Data Center (DMDC) in support of the DOD Personnel Identity Protection Program [2], and is meant to be a cost effective solution for securing access to applications across the ecosystem.

To date, there are roughly 3.5 million DSL credentials in circulation, securing around 20 different DOD and Department of Veterans Affairs applications. The system on a whole supports between 3 and 4 million authentications per month. Today, the DSL secures applications that manage data consisting of medical records, benefit applications, and banking information for active duty soldiers, family members, retirees, and veterans.

As usage of the credential grows, the potential for abuse against the DSL grows more attractive. For instance, automated systems, or even criminal enterprises, may seek access to the data protected by the credential. Thus, mechanisms to identify compromised credentials are increasingly important.

With permission from DMDC, and access to the DSL system artifacts, the goal of this thesis is to retrospectively inspect event patterns appearing in the DSL logs in order to detect potentially compromised credentials. Detecting these credentials is a step toward securing the DSL credential system and provides credential owners confidence that their personally identifiable information is secure. To achieve this goal, this thesis shall develop a set of procedures that leverage statistical analysis and industry standard tools to parse, store, index, and analyze the artifacts generated by the DSL system.

List of References

[1] W. E. Burr et al.. (2013). Electronic authentication guideline. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63-2.pdf.

[2] Department of Defense. (2004). Department of Defense personnel identity program. [Online]. Available: http://www.cac.mil/docs/DoDD-1000.25.pdf.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgements

Defense Manpower Data Center 400 Gigling Rd Seaside, CA 93955

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

From the moment a service member enlists in the military to the time of their passing, both the Department of Defense (DOD) and Department of Veterans Affairs (DVA) guarantee a variety of benefits for them and their family members. Benefits such as healthcare, housing, and continued education, are just a few of the rewards provided for their hardships.

These benefits are managed and maintained through a series of information technology (IT) systems that are susceptible to security vulnerabilities like any other modern day infrastructure. Criminal enterprises and individual scam artists look to capitalize on these vulnerabilities, for instance by attempting to falsely claim benefits for themselves. One of the main IT systems protecting these applications is the DOD Self Service Logon (DSL).

The DSL is a large-scale single-signon system, that issues username and password based credentials that adhere to the standards established in the National Institute of Standards and Technology (NIST) *Electronic Authentication Guidelines special publication (SP)* 800-63-2 [2]. It is used by individuals associated with the DOD and / or the DVA, such as active duty soldiers, family members, retirees and veterans, to access a wide range of self-service applications. On November 8 2010, the Under Secretary of Defense for Personnel and Readiness issued a memorandum requiring "All newly accessed Active Duty and National Guard and Reserve members in possession of a Common Access Card (CAC), shall be directed to obtain a DSL" [3]. Consequently, to date there are roughly 3.5 million DSL credentials in circulation, that are used to authenticate individuals to approximately 20 different applications. On a whole, the system supports between 3 and 4 million authentications per month and these numbers have consistently grown year after year. As shown in Figure 1, which we generate from internal Defense Manpower Data Center (DMDC) data, we can reasonably expect the growth of the system to continue in the near future making the need for this thesis more and more important.

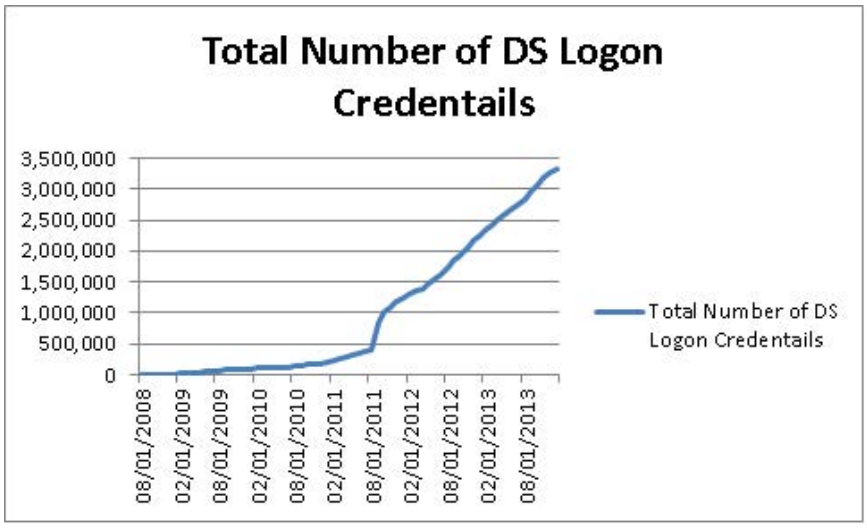

Figure 1.1: Growth of the DS Logon Credential

1.1 Background

Benefits that exist today within the DOD and DVA were created over time through various acts of the United States (U.S.) Congress. For example, the Dependents Medical Care Act of 1956 [4] and the Military Medical Benefits Amendments of 1966 [5] created programs that eventually became the military health system that exists today.

When a new benefit comes into existence a mechanism is needed that provides beneficiaries the ability to apply for and obtain that benefit. Often the DOD and DVA pursued a strategy to build a new IT system for each benefit. Over time this resulted in a patchwork of IT systems, that do not adhere to a cohesive architecture. As a result, each system implemented their own credentialing and authentication mechanism requiring beneficiaries to maintain countless usernames and passwords. To solve this problem both the DOD and DVA quickly recognized the need to implement a centralized credentialing and authentication mechanism to ease this burden and support these various systems.

In July, 2004, the DOD issued directive 1000.25 titled "DOD Personnel Identity Protection (PIP) Program" [6]. This program consolidated several identity programs from within the department, and centralized the capability to store authoritative identity information and issue DOD identity credentials. As outlined by this directive, the Defense Enrollment and Eligibility Reporting System (DEERS) repository became the authoritative source

for identity information used by the Real-time Automated Personnel Identification System (RAPIDS) to issue CACs to all DOD personnel, including military members, civilians, and contractors.

The CAC standard, later formalized by NIST in Federal Information Processing Standards (FIPS) 201 [7] and mandated across the entire federal government with Homeland Security Presidential Directive (HSPD) 12 [8], was an ideal solution for the authentication and authorization of DOD personnel to both physical and virtual DOD assets. Physical assets, such as base access and commissary privileges, and virtual assets, such as classified and non-classified networks, all became secured by the CAC infrastructure.

For authentication purposes, the CAC is considered a NIST level four credential, referring to the credential levels of assurance established in the NIST standards document SP 800-63-2. As shown in Table 1.1, there are four standardized NIST credential levels of assurance. This standard states, "Level 4 authentication is based on proof of possession of a key through a cryptographic protocol. At this level, in-person identity proofing is required" [2]. In terms of the DOD, this means that, in order to obtain a CAC, a person must prove his or her identity in-person at a RAPIDS station.

Table 1.1: NIST Credential Levels of Assurance, after [2].

Assurance Level	Description
1	Although there is no identity proofing requirement at this level, the authentication mechanism provides some assurance that the same Claimant who participated in previous transactions is accessing the protected transaction or data.
2	Provides single factor remote network authentication. At Level 2, identity proofing requirements are introduced, requiring presentation of identifying materials or information. A wide range of available authentication technologies can be employed at Level 2

3	Provides multi-factor remote network authentication. At least two authentication factors are required. At this level, identity proofing procedures require verification of identifying materials and information. Level 3 authentication is based on proof of possession of the allowed types of tokens through a cryptographic protocol.
4	Is intended to provide the highest practical remote network authentication assurance. Level 4 authentication is based on proof of possession of a key through a cryptographic protocol. At this level, in-person identity proofing is required.

While the CAC is ideal, and cryptographically secure for authentication purposes, the DOD realized not all of their assets required the same level of security provided by the CAC. They also realized it was cost prohibitive to issue CACs to family member, retiree, and veteran populations when those populations require access to lower-value assets. Therefore, the DOD needed an alternative to the CAC that was both secure and low cost.

In 2008, the DOD with support from DMDC, developed the DSL credential to be a NIST level two credential used for authentication purposes on self-service applications. Self-service applications are defined as, applications where a user manages their own information. For example, applications where users register for benefits, check the status of a medical claim, or update their contact information are all considered self-service. While self-service applications by definition do not maintain classified information, they can maintain sensitive beneficiary data (e.g., bank account numbers or medical history, such as Human Immunodeficiency Virus (HIV) status). Therefore, the DSL credential strictly adheres to security and identity proofing standards to ensure personally identifiable information (PII) remains protected and accessible only to the rightful owner. Had these standards not been enforced the DOD would have been liable for all data breaches and susceptible to fraud, waste, or abuse.

1.2 Goal

Due to the fact that the DSL system is a centralized credentialing and authentication single-signon system, if a credential is compromised all applications accepting a DSL are accessible. Consequently, as more and more applications throughout the DOD ecosystem migrate in support of the credential as shown in Appendix C, and with its continued growth as seen in Figure 1, the need to validate the credential's security has never been greater. Therefore, the security objective of this thesis is to help ensure our soldiers' hard earned benefits remain their hard earned benefits.

With permission from DMDC, and access to the DSL system artifacts, the goal of this thesis is to retrospectively inspect event patterns appearing in the DSL logs in order to detect potentially compromised credentials. Detecting these credentials is a step toward securing the DSL credential system and provides credential owners confidence that their PII is secure. To achieve this goal, this thesis shall develop a set of procedures that leverage statistical analysis and industry standard tools to parse, store, index and analyze the artifacts generated by the DSL system.

1.3 Contributions and Benefits

This thesis makes the following primary contributions, which we hope DMDC will leverage to enhance the security of the DSL system:

1. **Identified 84,712 Previously Unknown Potentially Compromised Credentials**
   As analyzed in Chapter 5, this thesis has flagged 84,712 potentially compromised credentials. These credentials were previously unknown to be suspect.
2. **Identify DSL Security Vulnerabilities**
   The results of this thesis help identify how the system is being exploited and shall lead to additional security measures to prevent future attacks (see §3.1.1).
3. **Establish Procedures for Handling Compromised Credentials**
   The results of this thesis shall lead to the development of standard operating procedures (SOPs) to follow after a credential is flagged as compromised. Steps such as, locking a credential, rolling back data modifications and potentially contacting the individual may all be part of the newly created procedures.

The remainder of this thesis is organized as follows:

- Chapter 2 discusses related work and the tools used to complete the analysis.
- Chapter 3 speaks to the artifacts provided by DMDC and any shortcomings that may exist within those artifacts.
- Chapter 4 reviews the steps required to process the artifacts provided and outlines a strategy used to identify potentially compromised credentials.
- Chapter 5 examines the results of the analysis conducted and determines if compromised credentials may exist.
- Chapter 6 explores the ground truth of credentials previously identified as potentially compromised.
- Chapter 7 summarizes the results, discusses future work.

CHAPTER 2:
Related Work and Tools

The work related to the identification of compromised credentials is found in three areas (i) big data and data visualization, (ii) automated network traffic and (iii) attack vectors of a single-signon system.

2.1 Big Data and Data Visualization

Data mining large datasets consisting of both structured and unstructured data has been a problem in the technology industry for a number of years. In the realm of unstructured data, Carlo Strozzi first used the term "No Structured Query Language (NoSQL)" in 1998 to "name his lightweight, open-source relational database that did not expose the standard Structured Query Language (SQL) interface" [9]. Over the years, the meaning of the term NoSQL has morphed to represent a classification of data storage systems that do not store data in a structured Relational Database Management System (RDBMS) but rather in a schema-free repository that is designed to simplify the storage and retrieval process.

The popularity of NoSQL repositories is mainly due to the fact that they address the data storage scalability problem while still maintaining performance by leveraging parallel processing. "NoSQL data stores will not be a "passing fad". The simplicity, flexibility, and scalability of these systems fills a market niche, (e.g., for web sites with millions of read-/write users and relatively simple data schemas)" [10].

The evolution of NoSQL repositories has lead to the need for tools, such as Splunk [11] and Elasticsearch [12], to replace the search capabilities previously provided by SQL queries. Elasticsearch, for example, provides the ability to perform full text searches on large, schema-free, unstructured repositories and provides key data visualization as needed.

In the work "Bridging the Gaps: Joining Information Sources with Splunk," Stearley *et al.* discuss the concept of indexing vast amounts of system log artifacts produced by the Red Sky supercomputer located at the Sandia National Laboratories [13]. They are able to identify critical issues within the supercomputer in almost real-time and produce statisti-

cally relevant graphs. This approach is not novel, but does provide a significant contribution in the development of our strategy on how to index the DSL artifacts provided by DMDC and potentially find compromised credentials.

Another key component of our strategy to finding compromised credentials is recognizing the artifacts provided by DMDC allow for time series analysis and the leveraging of data visualization techniques. "The aim of time series analysis is to obtain insight into phenomena, to discover repetitive patterns and trends, and to predict the future" [14]. A key contribution discussed in the work, "Strategies for the visualization of geographic time-series data" [15] is that a graphic can quickly be overloaded with labels and symbols and clustering strategies are needed to alleviate this issue. We leverage this insight in the initial phase of our analysis when searching for events from malicious countries (see Characteristic 4.2.1).

2.2 Automated Network Traffic

The strategies discussed in the works "Detecting Stealthy, Distributed SSH Brute-Forcing" [16] and "Classification of Automated Web Traffic" [17] provide concrete solutions on how best to identify automated traffic and ultimately compromised credentials. Javed and Paxson use descriptive statistics to establish a realistic threshold on the number of failed authentication attempts an average user may create within an hour when attempting to authenticate via Secure Shell (SSH) to a host. Using this threshold, they are able to identify system outliers by generating several Cumulative Distribution Functions (CDFs) from data stored within a host's syslogs. A critical component of their work relied on the ability to refine their threshold values by using a correlation data set. Their correlation data set consisted of known credentials attempting to brute-force the SSH system. Unfortunately, a correlation data set is not available in the case of the DSL, however, this approach is still valid and one that is heavily leveraged throughout this thesis.

While Javed and Paxson [16] used failed authentication attempts as one characteristic to detect brute-force attacks, the DSL has the opportunity to use several characteristics. Buehrer, Stokes, Chellapilla, and Platt identify several by looking at the different features of automated web traffic within the context of a search engine [17]. While the problem they are trying to solve is different than ours, the characteristics they use are applicable to this re-

search such as, “Number of Internet Protocols (IPs) / location,” “Reputation,” and “Request Time Periodicity.” All of which shall be explored in depth throughout this thesis.

Other works such as, “Automated Traffic Classification and Application Identification using Machine Learning” [18] attempt to classify automated network traffic by examining packet features. These features are examined with an unsupervised machine learning based system. While this is an effective approach as they are able to achieve an average accuracy of 86.5 percent [18]. Unfortunately, the required data to accomplish this technique is not available for our research. The artifacts provided by DMDC only supply the source IP for each event.

2.3 Attack Vectors of a Single-Signon System

Systems similar to the DSL have been developed by other organizations to maintain customer loyalty and simplify the authentication process for application providers. Several popular implementations are listed below.

- Single Sign-On with SAML on Force.com [19]
- SAML Single Sign-On (SSO) Service for Google Apps [20]
- Facebook Login [21]
- Claims-Based Single Sign-On for the Web and Windows Azure [22]

Most of these systems, however, authenticate NIST level one (see Table 1.1) credentials for any application provider. Where as the DSL system is a closed system authenticating NIST level two credentials for known application providers. While these systems differ they do share the same risks and attack vectors. Understanding the known attack vectors of these systems as discussed in the work “Risks of the passport single sign-on protocol” [23] is critical to identifying characteristics of compromised credentials. For example, Kormann’s description of an active attack, where an attacker acts as a proxy between the authentication system and the client, validated the decision to include the “Events from an IP that has a Poor Reputation” characteristic described in 4.2.1.

2.4 Tools

The related work discussed above starting with §2.1 provides foundation that which this thesis can build a strategy on how to identify potentially compromised credentials. To implement that strategy and develop a mechanism that is sustainable by DMDC a set of stable, robust tools are needed. While the following are not business requirements from DMDC adhering to these requirements when selecting tools shall make it easier for DMDC to continue this effort.

1. Zero Licensing Costs
2. Open Source
3. Ease of Deployment
4. Scalable
5. Large Support Community
6. Extensive Documentation

After performing a careful survey on the state of the art, three products were identified as potential candidates to support this thesis, Elasticsearch [12], Apache Solr [24] and Splunk [11]. While Splunk is a feature rich solution it can be immediately eliminated from the list due to the fact that it is a commercial closed source product. Even though, Elasticsearch and Apache Solr are similar, Elasticsearch has an extensive user community and documentation. Therefore, we use Elasticsearch for this thesis.

2.4.1 Elasticsearch

The Elasticsearch website defines their product as "A flexible and powerful open source real-time search analytics engine. Architected from the ground up for use in distributed environments where reliability and scalability are must haves, Elasticsearch gives you the ability easily to move beyond simple full-text search" [25]. In fact Elasticsearch is built on top of the Apache Lucene project which itself is a very mature open source Java based indexing and search technology. Elasticsearch is easily scalable by simply adding more machines and is extensible with a wide variety of company and community supported plugins that facilitate an ever growing list of uses for the tool. This thesis shall leverage Elasticsearch as its main indexing and search technology for all data.

Kibana

Data indexed and stored in Elasticsearch is searchable through an extensive Representational State Transfer (REST) Application Programming Interface (API) [26]; however, this interface makes visualizing valuable information challenging. To provide visualization, Elasticsearch built and maintains a plugin named "Kibana" to quickly and easily visualize data indexed with Elasticsearch. As described by Elasticsearch, "Kibana is Elasticsearch's data visualization engine, allowing you to natively interact with all your data in Elasticsearch via custom dashboards" [25]. Kibana is a web-based, javascript engine that allows for the filtering and visualization of data stored in Elasticsearch in many different formats e.g., maps depicting the geo-location or histograms for a time series. This thesis shall use Kibana to develop a series of dashboards intended to quickly identify compromised credentials.

2.4.2 MaxMind GeoLite

Every event within the artifacts provided by DMDC contains a source IP address. While IP addresses are only Internet identifiers, they can often be traced to a physical location in the real world. In fact there are several services that exist to perform this functionality.

The MaxMind Geo Lite service [27] is a free geo-location service that is used to find the physical location of an IP address. This thesis shall use the MaxMind GeoLite Java based API to identify the geo-location of every routable event within the artifacts provided by DMDC. This information is vital to being able to identify suspicious events and ultimately compromised credentials.

2.4.3 Spamhaus Zen

Along with a geo-location, every IP address has what is referred to as a reputation. IP reputations are generally maintained to assist network and system administrators in separating legitimate from illegitimate network traffic. While there are several commercial services that provide IP reputations, as with most things on the Internet free to use with community support is most often the best choice.

> The Spamhaus Project, "Is an international nonprofit organization whose mission is to track the Internet's spam operations and sources, to provide depend-

able realtime anti-spam protection for Internet networks, to work with Law Enforcement Agencies to identify and pursue spam and malware gangs worldwide, and to lobby governments for effective anti-spam legislation." [28]

This thesis shall use the Spamhaus Zen Domain Name System (DNS) Block List (DNSBL) service [29] to obtain the IP reputation for every event. As shown in Table 2.1 the Zen service has four categories of responses, Unsolicited, Snowshoe, Trap, and Policy. To ensure all events are assigned a reputation, good or bad, this thesis shall also assign two additional reputations, Allowlist, and Private. Accurate reputation information is vital in being able to identify suspicious events and corresponding compromised credentials.

Table 2.1: Spamhaus Zen Return Codes and Reputation Types

Type	Return Codes	Description
Unsolicited	127.0.0.2	Direct Unsolicited Bulk Email (UBE) sources, spam operations and spam services.
Snowshoe	127.0.0.3	Direct snowshoe spam sources detected via automation. "Snowshoe spamming is a technique used by spammers to spread spam output across many IPs and domains, in order to dilute reputation metrics and evade filters" [30].
Trap	127.0.0.4-7	The Composite Blocking List (CBL) is a list of suspected E-mail spam sending computer infections. The Composite Blocking List (CBL) takes its source data from very large spamtraps/mail infrastructures, and only lists IPs exhibiting characteristics such as, open proxies, worms/viruses/botnets , or are otherwise participating in a botnet and finally trojan horses or "stealth" spamware.

Policy	127.0.0.10-11	The Spamhaus Policy Block List (PBL) is a DNSBL database of end-user IP address ranges which should not be delivering unauthenticated Simple Mail Transfer Protocol (SMTP) email to any Internet mail server except those provided for specifically by an Internet Service Provider (ISP) for that customer's use.
Allowlist	NONE	IP does not appear on any block list.
Private	NONE	Refer to Request For Comment (RFC) 1918 10/8 prefix, 172.16/12 prefix, 192.168/16 prefix

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3:
System Artifacts

In many large organizations, the management of artifacts produced by its systems is a complex and challenging task. One that requires a deep understanding of data warehousing strategies in order to support the storage and recall of the artifacts. These artifacts often include structured data such as databases, eXtensible Markup Language (XML) and unstructured data such as system and application log files. This data can provide valuable information about an organization if strategies are available to process large quantities of information.

The DSL system is one of many systems developed and hosted by DMDC that produces all types of system artifacts. Within these artifacts lie key information about how and when a DSL credential is used. The main artifacts analyzed by this thesis are the DSL event logs. The event data is considered structured data and is currently stored in an Oracle RDBMS.

3.1 Events

When the DSL system was architected the auditing of a credential's usage was a first-order requirement – at all times it should be possible to produce a complete list of when, where, and how a credential is used. The result of this requirement was the event model, an API that is used throughout the system to capture and store "events" as they occur. The model is extensible so that over time, as new features are developed, new types of "events" can be recorded and audited. Please refer to Appendix A to see a complete listing of all the types of events captured by the system.

For the purposes of detecting compromised DSL credentials, this thesis shall examine event data from January 2013 to January 2014. The data provided by DMDC resulted in 51 artifacts each roughly 1GB in size and containing 1 million events. The artifacts were produced using Oracle's Data Pump Export Utility [31].

Each line in any one artifact represents a single event, that occurred within the system. In order to trace events back to a credential, each event is associated with exactly one profile.

As shown in Figure 3.1, a profile has a one-to-one relationship with a person registered in the DEERS repository and has exactly one DSL credential. Therefore, in order to have a DSL a person must be affiliated with the DOD and/or the DVA and registered in DEERS.

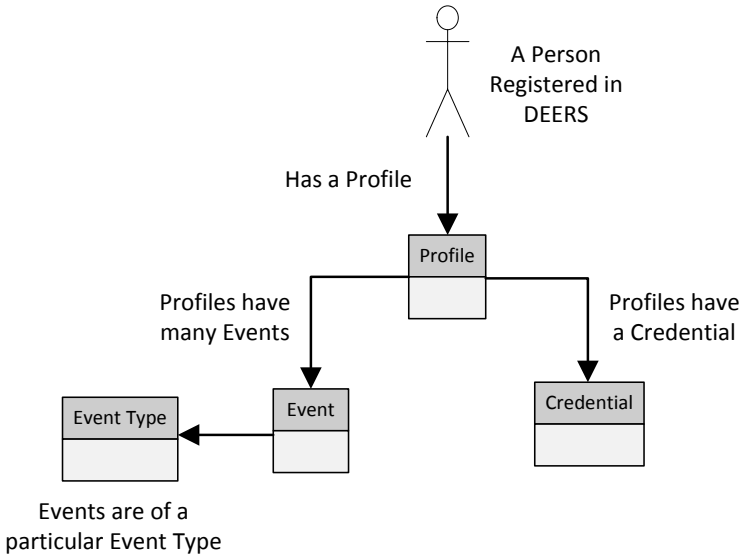

Figure 3.1: Entity Relationship Model

Understanding these relationships is critical to being able to discover compromised credentials as discussed in the analysis phase of this thesis (see Chapter 5).

3.1.1 Shortcomings

While extensive, the DSL event data has several shortcomings that are the direct result of policy and/or design decisions made by DMDC. As consumers of the data these shortcomings cannot be overcome at the time of our analysis.

The first shortcoming, is based on the fact that every event is tied to a profile. This means, before an event is saved, a user must first identify themselves. Users typically achieve this by supplying either PII (e.g., first and last name, date of birth, and social security number etc.) or a pre-registered DSL credential. Only when this information is provided, can the DSL system associate a user's traffic with a specific profile and create relevant events. Therefore, should an automated system attempt to compromise a credential by probing

the system with failed authentication attempts, this traffic is not recorded and cannot be detected or identified by our analysis because, the user was never properly identified.

The second shortcoming, is due to the fact that a subset of events contain IP addresses from the set of IPs defined in RFC 1918 [32]. Out of the 50,466,578 events provided by DMDC 6,709,884 or roughly 13.3 percent contain a RFC 1918 address. Prior to the event API going live in March 2012, all IP tracking and blocking was handled at the edge of the network and as a result no source IP data was transmitted internally. Consequently, until this issue was resolved by DMDC in March of 2013 all events tracked internal IP addresses. While this subset is small it does prevent us from capturing the geo-location of all events within the artifacts provided.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:
Methodology

Identifying patterns or specific data points when dealing with large amounts of data, structured or unstructured, is technically challenging. Even assuming sufficient physical architecture (e.g., storage, processing power, and memory) is already established to handle the vast amount of data, many challenges still exist. Items of interest are often obscured simply because there is too much data or too many complexities exist within the data to comprehend the relationships of how it all ties together.

Fortunately, many companies and organizations alike have developed strategies and tools to parse, store, index, and analyze massive amounts of data. By leveraging parallel processing and distributed computing strategies, the speed of analysis can scale sub-linearly with respect to the amount of data. These strategies have enabled organizations to make better decisions and be more efficient. We leverage these same strategies and tools to find those characteristics that are suggestive of compromised or otherwise anomalous credentials.

Our methodology includes a process to efficiently parse, store and index the artifacts provided by DMDC. Then we devise a set of characteristics that best describe compromised credentials and analyze the data to look for those characteristics. Based on combinations of characteristics that are outside of the norm, we identify potentially compromised credentials.

4.1 Process

A first and critical step of the process is to ensure the security of the artifacts provided by DMDC. Due to the sensitive nature of the artifacts and the fact that they contain PII (e.g., IP addresses), all data must be secured at all times, including when at rest. Therefore, we store and maintain the data on an isolated encrypted disk partition with access granted only on a need-to-know basis. Once we establish the encrypted partition, the installation of Elasticsearch and Kibana can begin.

4.1.1 Setup

The setup of the tools described in §2.4 is a relatively straight-forward process, which we accomplish with system administrator rights. Our analysis only requires a basic installation of the tools because, features such as, high availability are not necessary. A complete listing of all commands and configuration modifications needed to install the tools are listed in Appendix B.

4.1.2 Indexing

Upon completing the setup process, the indexing of the artifact data can commence. The strategy behind the process outlined below is based on a key understanding that Elasticsearch is an indexing and search engine that also stores JavaScript Object Notation (JSON) documents for fast retrieval. It is not designed to act like a typical RDBMS or NoSQL repository that balances resources between searches and modifications of existing records (e.g., INSERT and UPDATE). Therefore, the general strategy is to perform all modifications to event records prior to indexing in Elasticsearch, including all geo-location and IP reputation analysis. Our goal is that once the indexing is complete the analysis process shall be fast and easy.

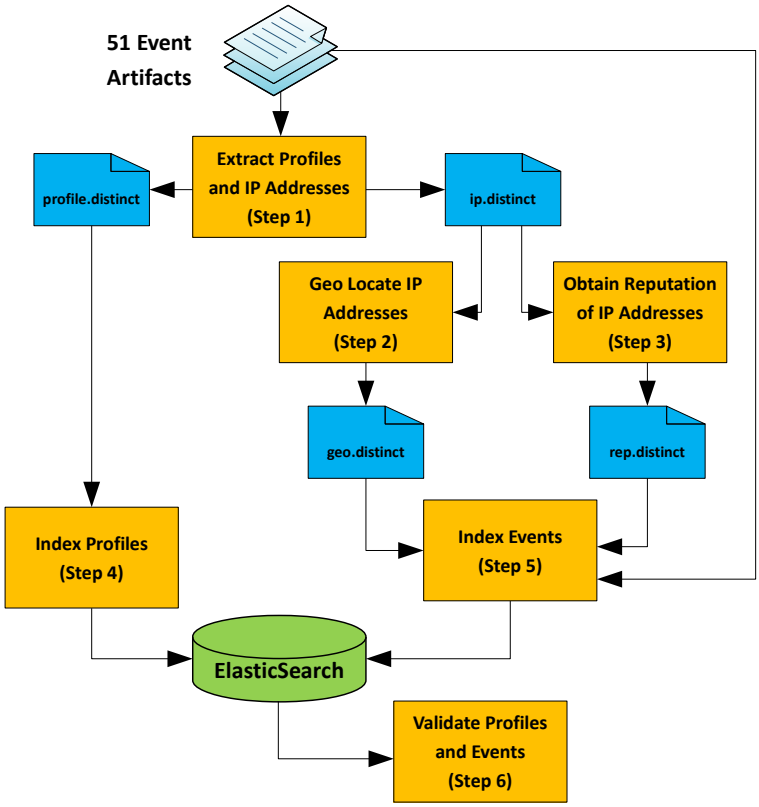

Figure 4.1: Indexing Process

As shown in Figure 4.1.2, there are six steps involved in our indexing process. Those steps are:

1. **Extract Profiles and IP Addresses**
   As discussed in §3.1 DMDC provided 51 artifacts each containing roughly 1 million events. For example, the following is a sample event (Note: White space has been removed to enhance readability and some values have been changed to ensure anonymity.)

   ```
   123456        33280374201301010000001490000    7127.0.0.1 IDM
   ```

   When parsed, we get a profile id of 123456 and an IP address of 127.0.0.1. After parsing roughly 50 million events we combine the results to produce two files,

"profile.distinct" and "ip.distinct", both of which are used in subsequent steps.

2. **Geo Locate IP Addresses**

   In order to obtain an IP address's geo-location in a efficient manner using the MaxMind Geo Lite service (see §2.4.2), the "ip.distinct" file produced in step one is broken down into several smaller files. Then using the MaxMind Geo Lite API we process each file and obtain a geo-location for each IP. The result of this step is a "geo.distinct" file that contains a list of comma delimited strings in the following format, "IP,LONGITUDE,LATITUDE".

3. **Obtain Reputation of IP Addresses**

   Using a similar process to the previous step, the "ip.distinct" file is once again broken down into several smaller files. This time, however, the reputation of each IP is obtained using the the Spamhaus Zen service (see §2.4.3). This step produces a "rep.distinct" file that contains a list of comma delimited strings in the format, "IP,REPUTATION+".

4. **Index Profiles**

   To store additional information about each profile (e.g., compromised characteristic see §4.2.1), we maintain the parent child relationship in Elasticsearch between the profile and the event data. Each profile is stored and indexed separate from the event data in the Elasticsearch repository by parsing the "profile.distinct" file and using the Elasticsearch Bulk API [33].

5. **Index Events**

   Similar to the previous step we store and index each of the roughly 50 million events using the Elasticsearch Bulk API [33]. The difference between this step and the previous one, is that prior to storing and indexing each event, we populate them with their corresponding geo-location and reputations previously identified in steps two and three.

6. **Validate Profiles and Events**

   The final step in the process is to validate that the previous steps accurately accounted for all profiles and events before the analysis can begin. We achieve this by obtaining a count from the "profile.distinct" file generated in step one and the 51 artifacts provided by DMDC. The "profile.distinct" file contains 3,687,780 profiles and each of the 51 artifacts except for one contained 1,000,000 events for a total of 50,466,578

events. Then by using the Elasticsearch Count API [34], we query the repository to validate these numbers match. Listing 4.1 shows the exact validation queries used to prove all profiles and events are properly indexed.

Listing 4.1: Actual Elasticsearch Validation Queries and Responses

```
curl http://localhost:9200/artifact/profile/_count
{
      "count": 3687780,
}
curl http://localhost:9200/artifact/event/_count
{
      "count": 50466578,
}
```

4.2 Strategy

The strategy of this thesis is to develop a list of characteristics that best describe a potentially compromised credential. Then, using the tools described above and descriptive statistics, such as CDFs, produce the set of credentials ordered by their likelihood of having been compromised. Finally, select a few of the identified credentials and attempt to find the "ground truth" by contacting the individuals via established channels such as the DMDC call center to determine if the credentials are truly compromised (see Chapter 6).

4.2.1 Characteristics of a Compromised Credential

Today, the DSL secures applications that manage data consisting of medical records, benefit applications and banking information for active duty soldiers, family members, retirees and veterans. In general, the accepting applications are designed to support human resources (HR) activities and the management of DOD or DVA benefits (see Appendix C for a complete list of the applications accepting a DSL). With these types of applications, one might expect that the typical user authenticates a couple of times per month to manage their benefits. In contrast, a profile with significantly more authentications is suggestive of a credential that is being abused. Naturally, a single characteristic is insufficient to accurately ascertain abuse. Therefore, we seek to assemble the multiple characteristics into a aggregate picture of the profile. The complete list characteristics are designed to identify

credentials that drastically exceed (fall within the top 0.5 percentile), their realistic estimates and are considered potentially compromised. The following is the complete list of characteristics we intend to examine.

Excessive Number of Events

Any credential with an excessive number of events, no matter the type (see Appendix A), suggests the credential is being used without human intervention. This characteristic is considered a catch all, in case the narrower scoped characteristics fail to identify excessive usage of a credential. For example, we do not have a characteristic looking for excessive password resets or registration events.

Excessive Number of Authentication Events

As described in §5.1.2, on average a typical user may authenticate to one of the DSL accepting applications once per month. Any credential with an excessive number of authentications may indicate a credential is compromised and is being used by an autonomous system.

Excessive Number of IP Addresses

In order to hide their activity and avoid detection, criminal enterprises frequently use a collection of proxy servers to carry out their attacks. According to Geer, “Attackers either can use their own computers to send bots and commands to victims or can use a machine they have infected, which then acts as a proxy server. These proxy servers can make finding the hacker difficult for security investigators" [35]. While it is normal for an average person to have several IPs accessing their credential, an excessive amount indicates criminal activity.

Excessive Number of Countries

Due to the fact the DSL population is comprised of DOD members and their families, there is a greater likelihood of a person accessing an application from overseas. Therefore, this characteristic is not designed to flag credentials accessing from overseas, but rather it is designed to flag credentials accessing from multiple countries. Once again this is a strategy for criminal enterprises to hide the location of their network traffic.

Excessive Number of Remote Proofing Failure Events

In order to obtain a NIST level two credential the DSL system employs a knowledge based identity verification system. This system asks a series of questions that are

very difficult to answer unless you are truly the individual. Any credential with an excessive number of remote proofing failures is an indication an attacker attempted to break the remote proofing process and compromise the credential.

Events from an IP that has a Poor Reputation

IP addresses accrue a reputation based on the activity that occurs at that address. Consequently, should any credential have an excessive number of events from an IP address with a negative reputation (e.g., Unsolicited, Snowshoe, or Trap) it is an indication the credential is compromised (see Table 2.1)

Events from Malicious Countries

If an event contains an IP that is geo-located to a country that is known to be a hot bed for criminal activity and/or is not an ally to the U.S. it may indicate the credential is compromised. Our baseline set of countries consider "malicious" is obtained from a *Bloomberg* 2013 article detailing the list of "Top 10 Hacking Countries" [36]. From this list we removed the U.S. and added Iran and North Korea for obvious geo-political reasons to make up a total of 11 "malicious" countries.

1. China (CN)
2. Russia (RU)
3. Turkey (TR)
4. Taiwan (TW)
5. Brazil (BR)
6. Romania (RO)
7. India (IN)
8. Italy (IT)
9. Hungry (HU)
10. Iran (IR)
11. North Korea (KP)

Periodicity of Authentication Events

People establish patterns in their daily lives. For example, most of us eat dinner at or around the same time every night. If we were to record these events and when they occurred over a period of time these patterns are easily identified. Using this same process, it is possible identify patterns of usage within a system and ultimately distinguish between a human or a machine. As a result, if an attacker has scripted the

use of a DSL credential to occur everyday at the exact same time these patterns are discoverable.

By calculating the Shannon Entropy [37] of a profile's authentication events we can identify those profiles that exhibit periodicity. We calculate the Shannon Entropy as part of this thesis using two different methods in order to identify all of the patterns within the system. To ensure enough data points exist to allow for randomness, both methods limit the profiles used in this characteristic to those with 100 or more authentication events.

In our first method, we calculate the entropy by ordering all authentication events by the time they occurred. Then we calculate the deltas in time between each event. We then input the delta values into our entropy calculation. With this method if a profile exhibits low entropy, or randomness, it is an indication the corresponding credential is potentially compromised.

Our second method, is very similar to our first, in the fact that we are only looking at the time each authentication event occurred. With this method, we extract the hour of the day (0 - 24) the event occurred. These values are then input into our entropy calculation, and any profile with a really low or high entropy is an indication the corresponding credential is potentially compromised.

4.2.2 Descriptive Statistics

Each of the eight characteristics discussed above in §4.2.1, produce a set of potentially compromised credentials. To find these sets, we use descriptive statistics to analyze the distribution of values for a characteristic across all profiles.

We calculate the minimum and maximum values to find the boundaries of our distribution and the mean, median and standard deviation values to describe the distribution. We also calculate the skewness and kurtosis to paint a general picture of the distribution. A positive skewness value, as shown in Figure 4.2.2, indicates the distribution tail falls to the right, where a negative skewness, also shown in Figure 4.2.2, indicates the tail falls to the left.

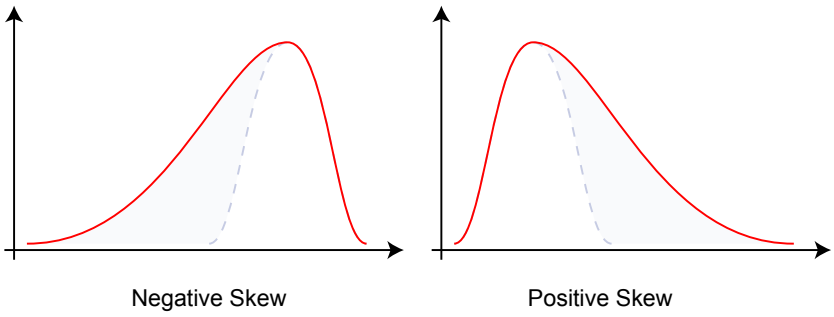

Figure 4.2: General Description of Skewness, from [1]

Similar to skewness, the kurtosis value describes the shape of the peak for a distribution. Generally a positive value indicates a majority of the values fall around the mean and the potential exists for a very long tail or extreme values. The descriptive statistics for each characteristic can be seen in the analysis Chapter 5.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5:
Analysis

After carefully establishing our infrastructure in §4.1.1 and indexing the artifacts provided by DMDC in §4.1.2, we implement the strategy discussed in §4.2, and commence our analysis to find potentially compromised credentials.

5.1 Results

Each of the eight characteristics (from §4.2.1) identify a set of potentially compromised credentials. We then combine these sets to produce a distinct set of profiles. Each profile correlates to a specific credential (see §3.1). We assign each profile a score based on weighted characteristic matches flagged for that profile. For example, if we flag profile 12345 for exhibiting an excessive number of events, and an excessive number of distinct IPs, it receives a score of two. Using this methodology we order the complete profile set by the score where profiles having the highest score have an increased likelihood of being compromised. The final results of our analysis are summarized in Chapter 7.

5.1.1 Excessive Number of Events

Out of the roughly 3.7 million profiles within the system, over the course of a year (see §3.1) the average profile contains 13 events. This value is significantly greater than the median value of 6 events, which indicates our distribution is tailed and contains statistical outliers that we want to find. We consider the 18,439 profiles within the top 0.5 percentile of the profiles with the most events to have an "Excessive" number of events. In fact, one credential has almost 34,000 events and is therefore an interesting candidate for further exploration, including ground truth. Table 5.1 and Figure 5.1.1 provide a complete description of our results for this characteristic.

Table 5.1: Excessive Number of Events Descriptive Statistics

Statistic	Value
Number of Profiles	3,687,780

Minimum Number of Events	1.0
Maximum Number of Events	33,993.0
Mean Number of Events	13.68
Standard Deviation	47.66
Median Number of Events	6.0
Skewness	188.76
Kurtosis	102,410.73
Number of Profiles over the estimate of realistic (1,000.0)	450.0
Number of Profiles in the Top 1 pct	37,183.0
Number of Profiles in the Top 0.5 pct	18,439.0
Number of Profiles in the Top 0.2 pct	7,382.0
Number of Profiles in the Top 0.1 pct	3,697.0

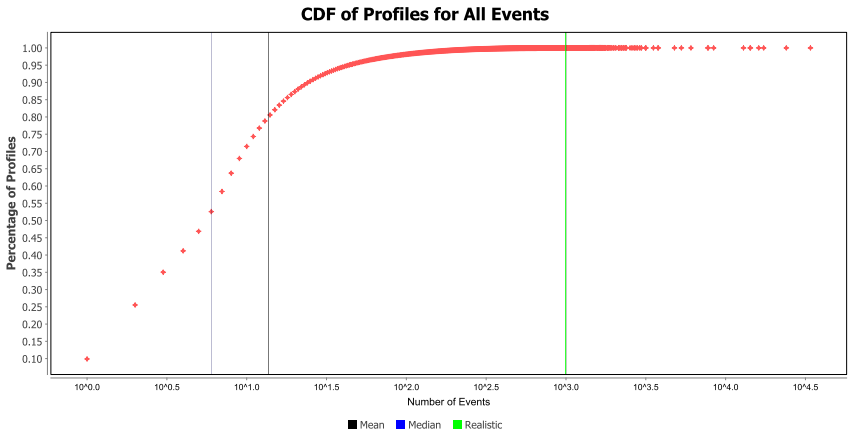

Figure 5.1: CDF of Profiles with Excessive Number of Events

5.1.2 Excessive Number of Authentication Events

Over the course of a year, each user on average authenticates once per month or 12 times a year. Interestingly, we discovered roughly 6,800 profiles authenticate on average at least once per day for a total of 365 authentications. This significantly breaks our a priori expectations of normal usage. In total, we find 15,643 profiles in the top 0.5 percentile and consider these to have "Excessive" authentication events. Table 5.2 and Figure 5.1.2 provide a complete description of our results for this characteristic.

Table 5.2: Excessive Number of Authentication Events Descriptive Statistics

Statistic	Value
Number of Profiles	3,115,280
Minimum Number of Events	1.0
Maximum Number of Events	33,989.0
Mean Number of Events	12.19
Standard Deviation	50.72
Median Number of Events	4.0
Skewness	185.15
Kurtosis	94,518.10
Number of Profiles over the estimate of realistic (365.0)	6796.0
Number of Profiles in the Top 1 pct	31,181.0
Number of Profiles in the Top 0.5 pct	15,643.0
Number of Profiles in the Top 0.2 pct	6,259.0
Number of Profiles in the Top 0.1 pct	3,129.0

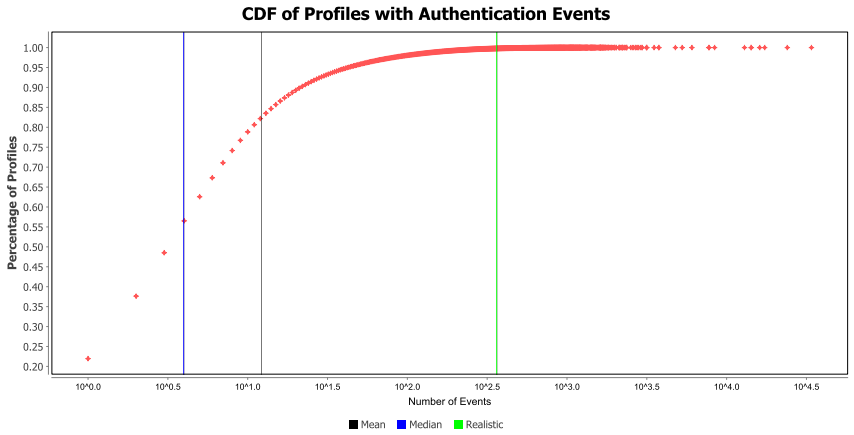

Figure 5.2: CDF of Profiles with Excessive Number of Authentication Events

5.1.3 Excessive Number of IPs Addresses

Out of the roughly 3.7 million profiles within the system, the average profile contains three different IP addresses. Over 120,000 profiles contain ten or more different IP addresses,

and the most extreme outlying profile has over 591 distinct IP addresses. There were 18,747 profiles in the top 0.5 percentile, which we consider "Excessive." Table 5.3 and Figure 5.1.3 provide a complete description of our results for this characteristic.

Table 5.3: Excessive Number of IP Addresses Descriptive Statistics

Statistic	Value
Number of Profiles	3,687,780
Minimum Number of IPs	1.0
Maximum Number of IPs	591.0
Mean Number of IPs	3.04
Standard Deviation	6.47
Median Number of IPs	2.0
Skewness	17.62
Kurtosis	586.61
Number of Profiles over the estimate of realistic (10.0)	122033.0
Number of Profiles in the Top 1 pct	40,054.0
Number of Profiles in the Top 0.5 pct	18,747.0
Number of Profiles in the Top 0.2 pct	7,414.0
Number of Profiles in the Top 0.1 pct	3,768.0

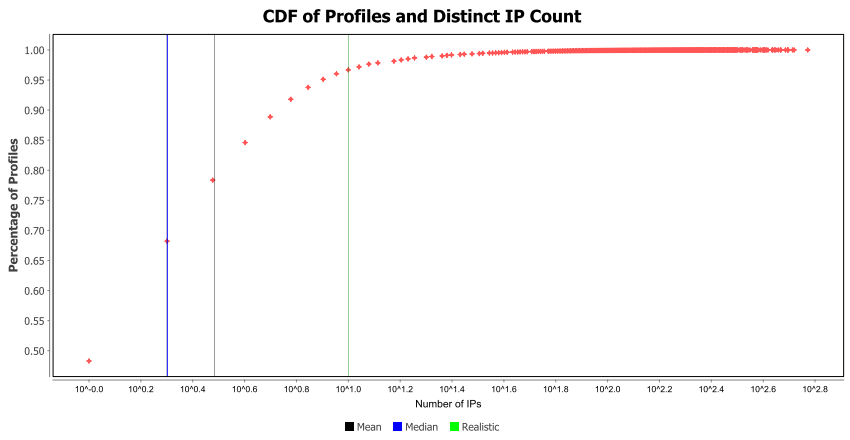

Figure 5.3: CDF of Profiles and Distinct IP Count

5.1.4 Excessive Number of Countries

The events with a geo-location (§3.1.1 discusses this and other shortcomings) show the average profile comes from a single country. With this particular characteristic, the top 0.5 percentile contains 50,882 profiles, which is interesting because, it is a number that is significantly greater than any other characteristic. More than likely it is the result of the fact that some of the DSL credential holders are in the DOD and often travel or live in foreign countries. Out of the 50,882 profiles only 496 exceeded our realistic estimate of three. Even though a majority of the profiles in the top 0.5 percentile fall below our estimate of realistic we are still going to flag them as "excessive" because, we use the top 0.5 percentile for the other characteristics. Table 5.4 and Figure 5.1.4 provide a complete description of our results for this characteristic.

Table 5.4: Excessive Number of Countries Descriptive Statistics

Statistic	Value
Number of Profiles	3,314,646
Minimum Number of Countries	1.0
Maximum Number of Countries	35.0
Mean Number of Countries	1.02
Standard Deviation	0.14
Median Number of Countries	1.0
Skewness	15.59
Kurtosis	1,412.59
Number of Profiles over the estimate of realistic (3.0)	496.0
Number of Profiles in the Top 1 pct	50,882.0
Number of Profiles in the Top 0.5 pct	50,882.0
Number of Profiles in the Top 0.2 pct	50,882.0
Number of Profiles in the Top 0.1 pct	50,882.0

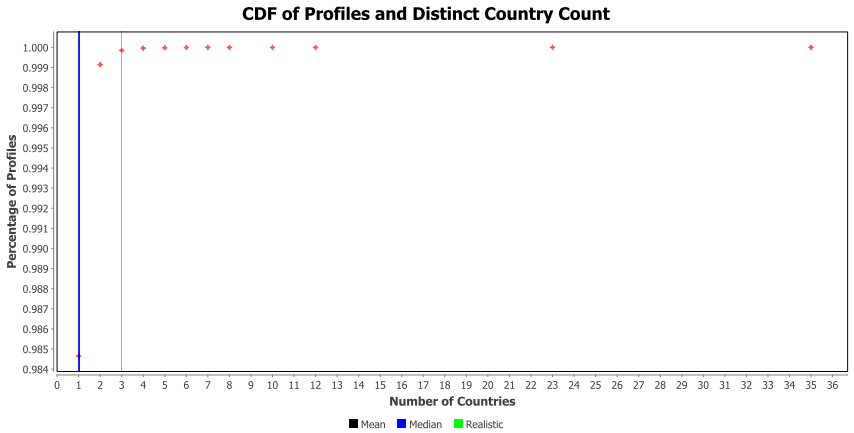

Figure 5.4: CDF of Profiles and Distinct Country Count

5.1.5 Excessive Number of Remote Proofing Failure Events

On average, users who fail a remote proofing, fail 1.8 times over a course of a year. Over 24,000 users fail the process more than six times, which exceeds our estimate of realistic. This results in 4,800 profiles in the top 0.5 percentile that we flag as "excessive." Table 5.5 and Figure 5.1.5 provide a complete description of our results for this characteristic.

Users who attempt to complete the process, but are never successful, may indicate an attacker is attempting to brute force their way through the remote proofing process. Future work, may want to investigate profiles with an excessive number of failures, but are eventually successful as a separate characteristic. This new characteristic would identify compromised credentials with a NIST level two identity proofing assurance.

Table 5.5: Excessive Number of Remote Proofing Failure Events Descriptive Statistics

Statistic	Value
Number of Profiles	766,777
Minimum Number of Events	1.0
Maximum Number of Events	338.0
Mean Number of Events	1.88
Standard Deviation	2.06
Median Number of Events	1.0

Skewness	14.65
Kurtosis	1,331.44
Number of Profiles over the estimate of realistic (6.0)	24,015.0
Number of Profiles in the Top 1 pct	8,329.0
Number of Profiles in the Top 0.5 pct	4,798.0
Number of Profiles in the Top 0.2 pct	1,941.0
Number of Profiles in the Top 0.1 pct	888.0

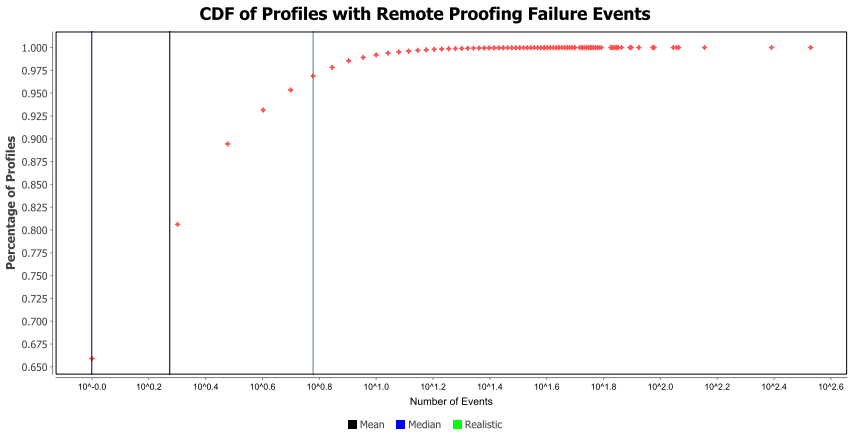

Figure 5.5: CDF of Profiles with Remote Proofing Failure Events

5.1.6 Events from an IP that has a Poor Reputation

The known abusive behaviors of compromised hosts include spam, attacks, and fraud. We therefore, investigate whether the set of IP addresses accessing a credential are known by external IP reputation systems. For example, a host that has sent an abusive email in the past may be a likely candidate to be used for other nefarious purposes. Unfortunately, the IP reputation data available to us for this thesis post dates the access events in our database. Therefore, due to the nature of the Internet and the shifting ownership of IP addresses only the top 0.5 percentile shall be considered. Of the 40,293 profiles containing events from an IP with a poor reputation only 203 are in the top 0.5 percentile. Table 5.6 and Figure 5.1.6 provide a complete description of our results for this characteristic.

Table 5.6: Events having Poor IP Reputation Descriptive Statistics

Statistic	Value
Number of Profiles	40,293
Minimum Number of Events	1.0
Maximum Number of Events	1,138.0
Mean Number of Events	5.39
Standard Deviation	15.35
Median Number of Events	2.0
Skewness	23.04
Kurtosis	1,054.77
Number of Profiles over the estimate of realistic (3.0)	14,671.0
Number of Profiles in the Top 1 pct	411.0
Number of Profiles in the Top 0.5 pct	203.0
Number of Profiles in the Top 0.2 pct	81.0
Number of Profiles in the Top 0.1 pct	42.0

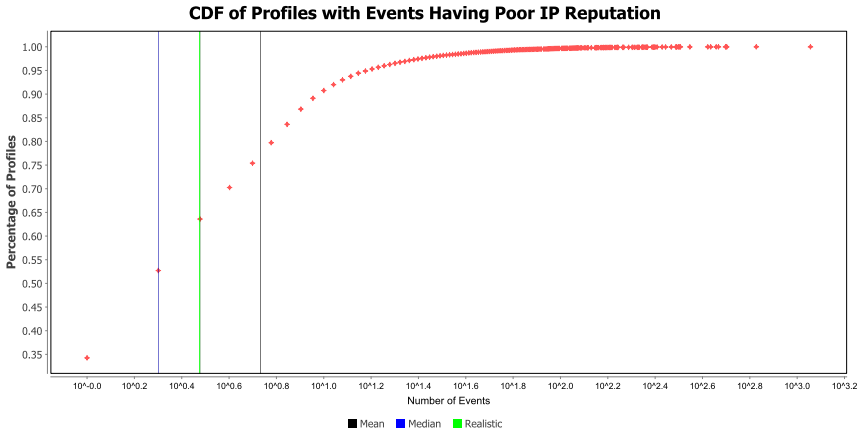

Figure 5.6: CDF of Profiles with Events having Poor IP Reputation

5.1.7 Events from Malicious Countries

Of the roughly 3.7 million profiles, only 4,469 are accessed from IP addresses that reside in a malicious country. One possible explanation for traffic originating from malicious

countries is the relative ease with which both legitimate and attacking users can mask their true geo-location behind a proxy server (see [35]). The end result is that only 23 profiles fall in the top 0.5 percentile to be flagged for additional investigation. Table 5.7 and Figure 5.1.7 provide a complete description of our results for this characteristic.

Table 5.7: Events from Malicious Countries Descriptive Statistics

Statistic	Value
Number of Profiles	4,469
Minimum Number of Events	1.0
Maximum Number of Events	372.0
Mean Number of Events	6.97
Standard Deviation	15.85
Median Number of Events	3.0
Skewness	10.68
Kurtosis	163.55
Number of Profiles over the estimate of realistic (10.0)	633.0
Number of Profiles in the Top 1 pct	46.0
Number of Profiles in the Top 0.5 pct	23.0
Number of Profiles in the Top 0.2 pct	9.0
Number of Profiles in the Top 0.1 pct	5.0

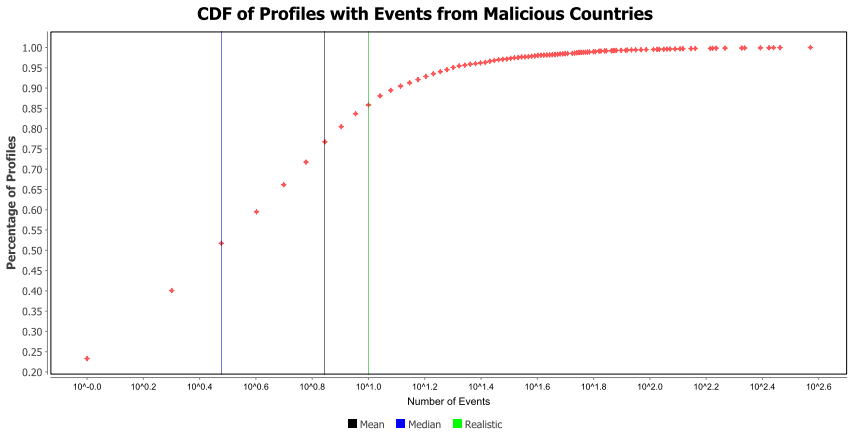

Figure 5.7: CDF of Profiles with Events from Malicious Countries

5.1.8 Periodicity of Authentication Events

As discussed in 4.2.1 we limit the number profiles when looking for periodicity to only those containing 100 or more authentication events. The effect of this decision allows us to focus on only 61,417 profiles when computing the entropy using both methods also described in 4.2.1.

With our first method, the delta approach, we discover 181 profiles with an entropy between 0 and 1. This characteristic is different than the seven others previously discussed because, in this case we are looking for the results with low entropy at the bottom of the result set. Consequently, there were 303 profiles in the bottom 0.5 pct that shall be flagged for further investigation. Table 5.8 and Figure 5.1.8 provide a complete description of our results for this method.

Table 5.8: Periodicity of Authentication Events Statistics

Statistic	Value
Number of Profiles	61,417
Minimum Entropy	0.0
Maximum Entropy	6.35
Mean Entropy	5.22
Standard Deviation	0.64
Median Entropy	5.35

Skewness	-2.55
Kurtosis	13.10
Number of Profiles in the Bottom 1 pct	614
Number of Profiles in the Bottom 0.5 pct	307
Number of Profiles in the Bottom 0.2 pct	123
Number of Profiles in the Bottom 0.1 pct	61

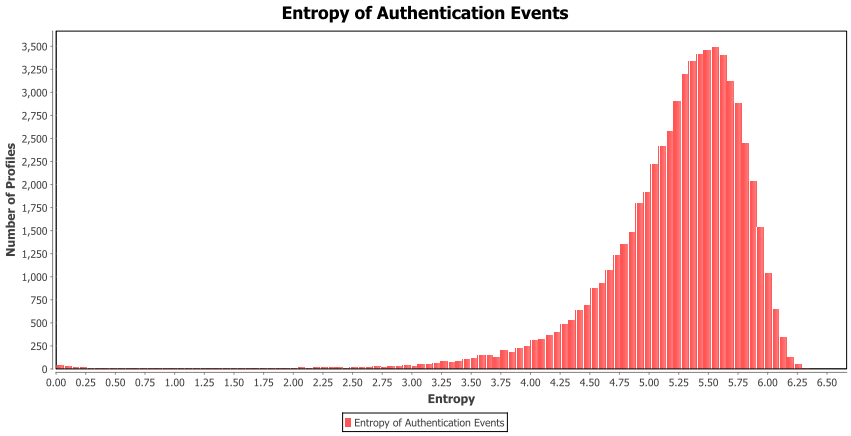

Figure 5.8: Histogram for Periodicity of Authentication Events

Our second method, the 24 hour bin approach, we only discover 98 profiles with an entropy between 0 and 1, however, with this method both low and high entropy values are important. A high entropy indicates authentications occurred at the exact same time of day. Consequently, we found 234 profiles with an entropy between 4.48 and 4.58, which is the max value possible. Table 5.9 and Figure 5.1.8 provides a complete description of our results for this method.

Table 5.9: Periodicity of Authentication Events with 24 Hour Bins Statistics

Statistic	Value
Number of Profiles	61,417
Minimum Entropy	0.0
Maximum Entropy	4.58

Mean Entropy	3.84
Standard Deviation	0.39
Median Entropy	3.92
Skewness	-2.00
Kurtosis	10.84
Number of Profiles in the Bottom 1 pct	614
Number of Profiles in the Bottom 0.5 pct	307
Number of Profiles in the Bottom 0.2 pct	123
Number of Profiles in the Bottom 0.1 pct	61

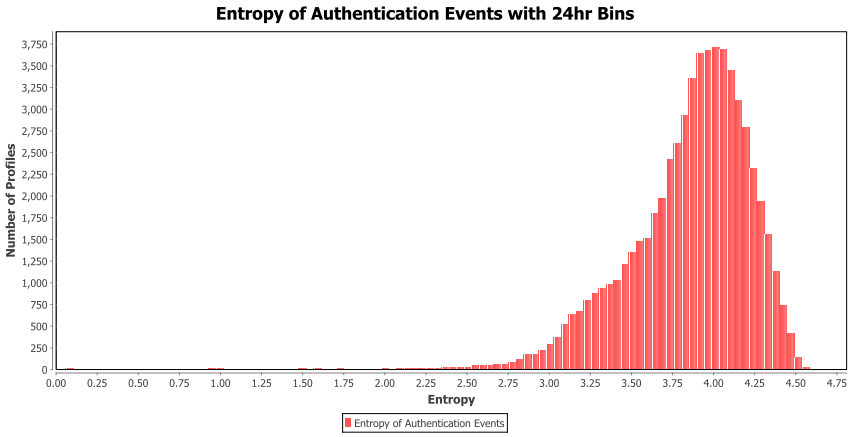

Figure 5.9: Histogram for Periodicity of Authentication Events With 24 Hour Bins

CHAPTER 6:
Ground Truth

To further investigate outliers identified in the preceding analysis, DMDC, the owner of the DSL system, agreed to contact the owners of credentials flagged as suspicious. DMDC maintained SOP when contacting individuals for further information. In an effort to maintain anonymity, credentials owners shall be referred to as profile "one," "two," and "three."

6.1 Profile One

As one of the 28 profiles having five out of the eight characteristics (see Table 7.1), obtaining ground truth for profile one is important. Upon contacting profile one, DMDC introduced themselves and verbally authenticated the individual's identity. DMDC proceeded to explain why they were contacting the individual and began to inquire if they have ever traveled to any of the countries associated with the activity seen in their DSL credential event logs (see Figure 6.1).

Figure 6.1: Geo Locations of Profile One Events (Produced by Kibana)

The individual responded they never visited these countries. However, profile one had an explanation for the unusual activity. The individual expressed a strong concern for Internet privacy, and, as a result, how they conduct all business using the Tails operating system. The Tails website states, "Tails is a live operating system, that you can start on almost any computer from a DVD, USB stick, or SD card. It aims at preserving your privacy and anonymity" [38]. Tails achieves this by requiring all Internet connections go through

the anonymizing Tor network. Originally announced in 2002, by Roger Dingledine, Nick Mathewson, and Paul Syverson and described in [39], Tor allows users to conceal their identity and hide their activity by encrypting packets and routing them through a series of Tor nodes hosted across the globe.

Understanding how profile one uses the DSL credential and accesses the Internet, explains most of the characteristics associated with their credential. While this is not an instance of fraud, waste or abuse, it does validate that our work is able to successfully identify credentials that are being used by individuals trying to hide their activity.

6.2 Profile Two

Profile two exhibits several characteristics of interest: excessive number of events, excessive number of authentication events, and high periodicity of authentication events. As shown in Figure 6.2, profile two over the course of a year has accumulated over 23,000 authentication events and has an entropy of 0.0033, which is low when compared to other profiles and is suggestive of an automated system performing authentications with this credential.

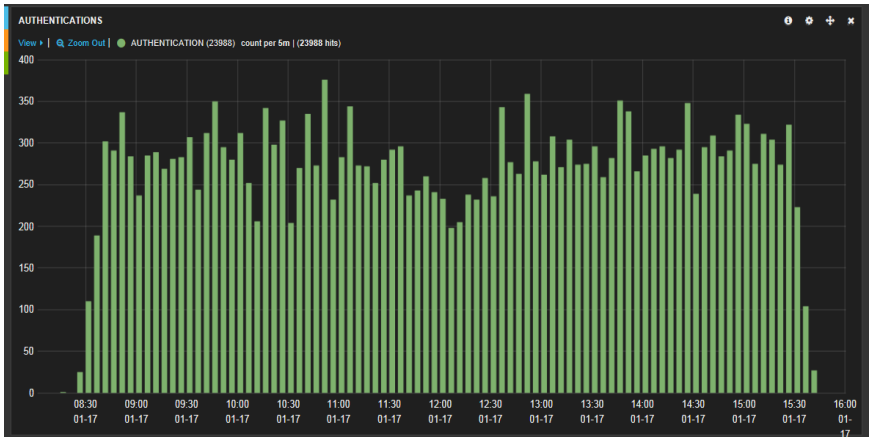

Figure 6.2: Authentication Events for Profile Two (Produced by Kibana)

After further research, prior to contacting this individual, DMDC discovered these authentications are attributable to an individual performing system load testing from a DMDC partner (see Appendix C) data center. While such load testing goes against DMDC policy, it does prove we are able to identify credentials used by an automated system.

6.3 Profile Three

Flagged due to failing an excessive number of remote proofing failures (see Figure 6.3, and exhibiting periodicity for authentication events, profile three was contacted by DMDC in an effort to find the ground truth. Upon contacting the individual, authenticating their identity, and explaining the situation the individual advised DMDC all authentications are legitimate.

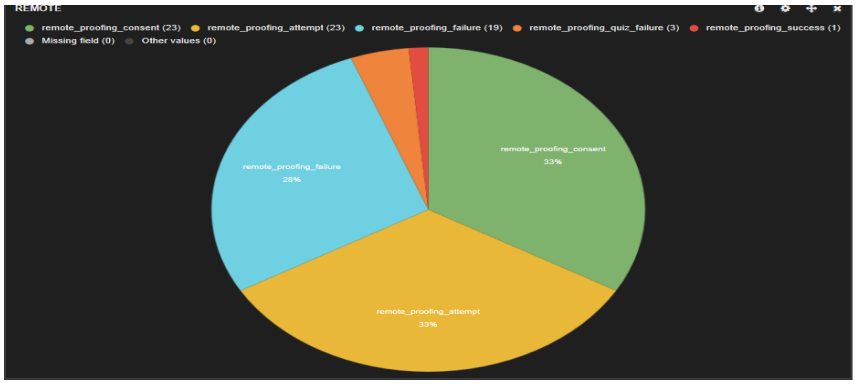

Figure 6.3: Remote Proofing Events for Profile Three (Produced by Kibana)

As a veteran of the U.S. military, they stated they have Attention Deficit Hyperactivity Disorder (ADHD), Post-Traumatic Stress Disorder (PTSD), anxiety, they are neurotic, and have memory issues. They engage in repetitive behaviors. As a result, they make frequent updates to their records because they are anxious about their information changing. They also check for any changes, new features, updates to their claim status, etc.. Their memory issues led them to not answer quiz questions correctly and they frequently forget their password.

The response provided by this individual, explains the associated characteristics for this credential. It is also a reality check of how hard it is to simply identify potentially compromised credentials based solely on a few characteristics when each human is so unique.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 7:
Conclusion

In this thesis, we posit that potentially compromised credentials are detectable by analyzing the system artifacts of a large-scale production, single-signon system. With permission from DMDC, and access to the DSL system artifacts, we parsed, stored, indexed, and analyzed the event logs representing the actions taken by roughly 3.5 million credentials between January 2013 to January 2014.

We pursued a strategy that identified eight characteristics compromised credentials may possess (see §4.2.1). We then used statistical analysis and a narrowing threshold (e.g., top 1.0 percentile, top 0.5 percentile etc.) to focus our results on the most extreme cases within each characteristic. Finally, we combined the results from each characteristic to produce a weighted set of potentially compromised credentials for each threshold category. Table 7.1 summarizes our results by showing the breakdown between the number of credentials and the number of characteristics they hold, within each threshold category. While Figure 7 illustrates how the number of credentials identified is reduced by narrowing the threshold used to flag an outlying credential.

Table 7.1: Weighted Compromised Credentials

No. of Characteristics	Top 1.0%	Top 0.5%	Top 0.2%	Top 0.1%
1	80,818	67,045	57,885	54,417
2	20,400	11,234	4,897	2,577
3	13,740	5,792	1,924	895
4	1,601	613	191	76
5	96	28	5	1
Total	116,656	84,712	64,902	57,966

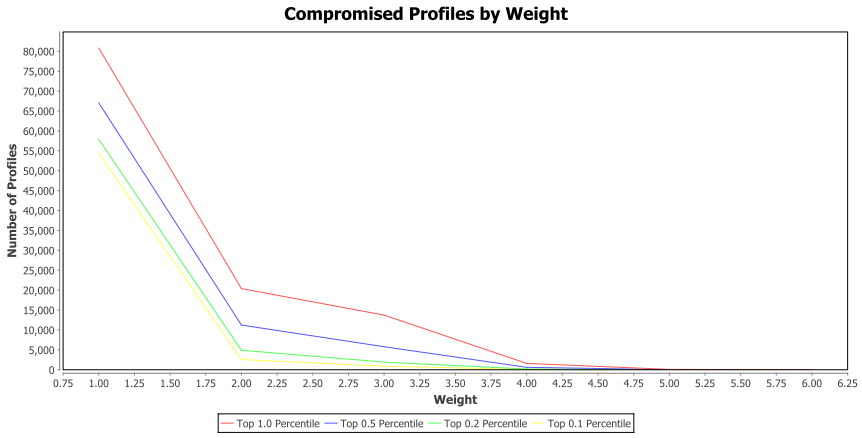

Figure 7.1: Weighted Compromised Credentials

We recognize the fact that our results contain both false negatives and false positives, as seen in our ground truth analysis (see Chapter 6). We also recognize, that we have not identified all compromised credentials within the DSL system. For example, there are many known instances reported to DMDC support centers of ex-spouses compromising credentials because, they are able to successfully complete the remote proofing process for their ex-partners. Our strategy (see §4.2), is not designed to flag profiles with these types of normal usage patterns. Therefore, credentials compromised in this manner are not identified by our research.

Having said that, our research has produced very valuable results for the DOD and DMDC. Take for example the top 0.5 percentile threshold, where we discovered 84,712, highly suspicious, potentially compromised credentials. Within this set, we identified 641 credentials that have four or more characteristics. It is our recommendation, that DMDC take immediate action to ensure the rightful owners are in control of these credentials. We recognize there is a cost, time and / or money, associated with completing this task. Therefore, we minimally recommend these credential holders be forced to reproof their identity and / or have the DMDC support center send an email to notify these individuals of the suspicious activity associated with their credential.

In conclusion, it is our belief the most valuable aspect of our research is the repeatable process we have developed. If this process is further refined, over time it has the potential to accurately and in real-time identify compromised credentials. Pursuing this effort along

with additional security enhancements, such as multiple factors of authentication, shall significantly increase the security footprint of the DSL credential. It is our hope, the DOD and DMDC continue to pursue this research to ensure the viability and sustainability of the DSL system.

7.1 Future Work

To improve the quality and quantity of credentials flagged as compromised, an in depth analysis evaluating the precision and recall of our approach is required. To accomplish this task, the DMDC support center could be used to determine if all credentials identified are truly compromised. As discussed above and as seen in our ground truth investigation (see Chapter 6), there is a cost associated with completing this analysis using this technique. However, this effort is required in order to increase the accuracy of credentials flagged as compromised.

In addition to this analysis, future work must develop additional characteristics in order to increase the accuracy of our approach. Taking our results as an example, of the 84,712 credentials we flag as compromised in the top 0.5 percentile, 641 are considered highly suspicious since they possess at least four of the eight characteristics. With additional characteristics, we can achieve greater accuracy and reduce the number of highly suspicious credentials. This assumes each additional characteristic is justifiable with sound logic. For example, of the 641 credentials identified as highly suspicious, how many are currently receiving some form of financial benefit from the U.S. government? Or, how many of these credentials are associated with deceased individuals not reported to the DOD? Using additional data sources and some level of creativity, increasing the accuracy of credentials flagged as highly suspicious is easily within the realm of possible.

Our research is reactive since we analyze the data long after the events actually occurred. Future work may easily address this issue and become more proactive by obtaining a live data feed of events from the DSL system that are piped into Elasticsearch. At that point, a Kibana dashboard could easily be developed to flag compromised credentials in real-time. Having the ability to detect compromised credentials in real-time is critical to prevent future criminal enterprises or individual scam artists from executing large scale attacks against the DSL system.

Our approach to using statistical analysis is not unique, however, we apply it to a new problem domain and believe the general approach is under-utilized and is readily applicable to other problem spaces. For example, a significant number of failed CAC authentications at a physical DOD location may indicate an individual attempting to gain illegal access to a facility. Or a physician, with an excessive number of claims may be an indication of fraud within the DOD or DVA healthcare system. These and other problem domains shall be left for future work and further research.

APPENDIX A:
Event Types

Table A.1: Table of Event Types tracked by the DSLs system

Code	Event	Summary
1	REGISTRATION	A new credential is registered and associated with an identity
2	ACTIVATION	A new credential is activated for use within the system.
3	DEACTIVATION	A credential has been deactivated and can no longer be used.
4	UPGRADE	The identity associated with a credential has been proof to a higher NIST level of assurance.
5	MAIL_SENT	A credential activation letter has been sent via United States Postal Service (USPS).
6	MAIL_RETURNED	A credential activation letter has been returned via USPS.
7	AUTHENTICATION	A successful authentication has occurred for the associated credential.
8	SURROGATE_ACCESS	A surrogate has accessed the credential owner's information.
9	USERNAME_RETRIEVED	The user name associated with the credential has been retrieved.
10	PASSWORD_RESET	The password associated with the credential has been reset.
11	CHALLENGE_QUESTIONS_RESET	The challenge questions associated with the credential have been reset.

12	PERSONAL_IMAGE_RESET	The security image used during authentication has been reset.
13	DEVICE_ACTIVATION	A second factor device has been registered / activated for use with the credential.
14	DEVICE_DEACTIVATION	A second factor device has been deregistered / deactivated for use with the credential.
15	RELATIONSHIP_ESTABLISHED	A surrogate relationship has been established.
16	PERMISSION_GRANTED	A new permission has been granted to a surrogate relationship.
17	PERMISSION_DENIED	A permission has been removed from a surrogate relationship.
18	PERSON_PULLED	An operator has requested to view the credential owner's information.
19	PERSON_ADDED	An operator has added a new person to the repository
20	PROFILE	A profile is created that represents a person's identity within the DSL system.
21	REMOTE_PROOFING_ATTEMPT	An attempt to remote proof a person's identity to increase the NIST level assurance has occurred.
22	REMOTE_PROOFING_SUCCESS	An attempt to remote proof a person's identity to increase the NIST level assurance was successful.
23	REMOTE_PROOFING_FAILURE	An attempt to remote proof a person's identity to increase the NIST level assurance has failed.

24	REMOTE_PROOFING_DISCOVERY_FAILURE	The remote proofing vendor failed to find a person's identity with the information provided.
25	REMOTE_PROOFING_VELOCITY_FAILURE	The remote proofing vendor failed to proof a person's identity due to too many successive failures.
26	REMOTE_PROOFING_VERIFICATION_FAILURE	The remote proofing vendor found a person's identity however, the information provided failed to satisfy the verification requirements.
27	REMOTE_PROOFING_QUIZ_FAILURE	The attempt to answer the remote proofing quiz failed.
28	REVOCATION	The credential has been revoked.
29	EMAIL_SENT	A digitally signed email has been sent to the credential owner's registered email address.
30	USERNAME_CHANGED	The user name associated with the credential has been change to the DOD Enterprise User Name.
31	REMOTE_PROOFING_CONSENT	A user has granted consent to remotely proof an identity.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B:
Elasticsearch Setup

```
# Download and install Java 1.7 JDK
rpm --install [JDK FILE]

# Elasticsearch is stored on the encrypted partition
cd /mnt/drobo/dmdc_encrypted/

# Download and install Elasticsearch from download.elasticsearch.org
mkdir elasticsearch
cd elasticsearch
wget https://.../elasticsearch-1.1.0.tar.gz
tar -xzvf elasticsearch-1.1.0.tar.gz
ln -s elasticsearch-1.1.0/ latest
cd latest

# Download and install Kibana from download.elasticsearch.org
cd /mnt/drobo/dmdc_encrypted/elasticsearch/
wget https:///.../kibana-3.0.0.tar.gz
tar -xzvf kibana-3.0.0.tar.gz
cd /mnt/drobo/dmdc_encrypted/elasticsearch/latest
mkdir plugins
cd plugins
mkdir kibana
ln -s /mnt/drobo/dmdc_encrypted/elasticsearch/kibana-3.0.0 _site

# Modify Elasticsearch config to separate data from installation
vi /mnt/drobo/dmdc_encrypted/elasticsearch/latest/config/elasticsearch.yml
# cluster.name: analysis-cluster-1
# path.data: /mnt/drobo/dmdc_encrypted/elasticsearch/data

# Set Owners and Permissions
```

```
chown -R elasticsearch:dmdc /mnt/drobo/dmdc_encrypted/elasticsearch
chmod -R 744 /mnt/drobo/dmdc_encrypted/elasticsearch

# Start Elasticsearch as the elasticsearch user
./bin/elasticsearch -Xmx4g -Xms3g -d
```

APPENDIX C:
Applications Accepting DS Logon

1. eBenefits
2. TRICARE Online (TOL)
3. milConnect
4. DS LOGON Self-Service website
5. Beneficiary Web Enrollment (BWE)
6. Address Update
7. Family Subsistence Supplemental Allowance (FSSA)
8. Joint Qualification System (JQS)
9. RAPIDS Self Service (RSS)
10. Transition GPS
11. DMDC Reserve Component Purchased TRICARE Application (RCPTA)
12. Verification of Military Experience and Training (VMET)
13. Health Net Federal Services
14. Humana Military
15. MetLife
16. myTRICARE
17. TRICARE Overseas
18. TRICARE4u
19. DoD Spouse Education and Career Opportunities (SECO)

THIS PAGE INTENTIONALLY LEFT BLANK

REFERENCES

[1] "Skewness," *Wikipedia*. [Online]. Available: http://en.wikipedia.org/wiki/Skewness, [Accessed May 17, 2014].

[2] W. E. Burr *et al.*, "Electronic authentication guideline," National Institute of Standards and Technology, Tech. Rep., 2013. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63-2.pdf.

[3] "Mandatory self-service logon required," August 2013. [Online]. Available: http://www.hanscom.af.mil/news/story_print.asp?id=123360514.

[4] "Dependents Medical Care Act. Pub L. No. 569." June 1956. [Online]. Available: http://www.gpo.gov/fdsys/pkg/STATUTE-70/pdf/STATUTE-70-Pg250.pdf.

[5] "Military Medical Benefits Amendments. Pub L. No. 89-614." September 1966. [Online]. Available: http://www.gpo.gov/fdsys/pkg/STATUTE-80/pdf/STATUTE-80-Pg862.pdf.

[6] "Department of Defense. (2004). Department of Defense personnel identity program." [Online]. Available: http://www.cac.mil/docs/DoDD-1000.25.pdf.

[7] Computer Security Division, "Personal identity verification (PIV) of federal employees and contractors," National Institute of Standards and Technology, Tech. Rep., August 2013. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.201-2.pdf.

[8] "Department of Homeland Security. (2004, August). Homeland Security Presidential Directive-12." [Online]. Available: https://www.dhs.gov/homeland-security-presidential-directive-12#1.

[9] "NoSQL," *Wikipedia*. [Online]. Available: http://en.wikipedia.org/wiki/NoSQL#History, [Accessed May 10, 2014].

[10] R. Cattell, "Scalable sql and nosql data stores," *ACM SIGMOD Record*, vol. 39, no. 4, pp. 12–27, 2011.

[11] "Splunk." [Online]. Available: http://www.splunk.com.

[12] "Elastic search." [Online]. Available: http://www.elasticsearch.org.

[13] J. Stearley *et al.*, "Bridging the gaps: Joining information sources with splunk," in *Proceedings of the 2010 Workshop on Managing Systems via Log Analysis and Machine Learning Techniques*, 2010, pp. 8–8.

[14] J. J. Van Wijk and E. R. Van Selow, "Cluster and calendar based visualization of time series data," in *Proceedings of the 1999 IEEE Symposium on Information Visualization*. IEEE, 1999, pp. 4–9.

[15] M. Monmonier, "Strategies for the visualization of geographic time-series data," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 27, no. 1, pp. 30–45, 1990.

[16] M. Javed and V. Paxson, "Detecting stealthy, distributed ssh brute-forcing," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*. ACM, 2013, pp. 85–96.

[17] G. Buehrer *et al.*, "Classification of automated web traffic," *Weaving Services and People on the World Wide Web*, 2009.

[18] S. Zander *et al.*, "Automated traffic classification and application identification using machine learning," in *The IEEE Conference on Local Computer Networks, 2005. 30th Anniversary.* IEEE, 2005, pp. 250–257.

[19] "Single sign-on with saml on force.com." [Online]. Available: https://developer.salesforce.com/page/Single_Sign-On_with_SAML_on_Force.com.

[20] "Saml single sign-on (sso) service for google apps." [Online]. Available: https://developers.google.com/google-apps/sso/saml_reference_implementation.

[21] "Facebook login." [Online]. Available: https://developers.facebook.com/docs/facebook-login/v2.0.

[22] "Claims-based single sign-on for the web and windows azure." [Online]. Available: http://msdn.microsoft.com/en-us/library/ff359102.aspx.

[23] D. P. Kormann and A. D. Rubin, "Risks of the passport single signon protocol," *Computer Networks*, vol. 33, no. 1, pp. 51–58, 2000.

[24] "Apache solr." [Online]. Available: http://lucene.apache.org/solr.

[25] "Overview." [Online]. Available: http://www.elasticsearch.org/overview/.

[26] "Rest api." [Online]. Available: http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/docs.html.

[27] "Maxmind geo lite." [Online]. Available: http://www.maxmind.com.

[28] "Spamhaus organization." [Online]. Available: http://www.spamhaus.org/organization/.

[29] “Spamhaus zen.” [Online]. Available: http://www.spamhaus.org/zen/.

[30] “Snowshoe spamming.” [Online]. Available: http://www.spamhaus.org/faq/section/Glossary#233.

[31] “Oracle database utilities 11g release 1 (11.1): Data pump export,” 2011. [Online]. Available: http://docs.oracle.com/cd/B28359_01/server.111/b28319/dp_export.htm#i1007851.

[32] Y. Rekhter *et al.*, “Address allocation for private internets,” Network Working Group, memorandum, Internet Engineering Task Force, Feb 1996. [Online]. Available: http://www.ietf.org/rfc/rfc1918.txt.

[33] “Bulk api.” [Online]. Available: http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/docs-bulk.html.

[34] “Count api.” [Online]. Available: http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/search-count.html.

[35] D. Geer, “Malicious bots threaten network security,” *Computer*, vol. 38, no. 1, pp. 18–20, 2005.

[36] M. Milian, “Top ten hacking countries,” *Bloomberg*, April 2013. [Online]. Available: http://www.bloomberg.com/slideshow/2013-04-23/top-ten-hacking-countries.html.

[37] “Shannon Entropy,” *Wikipedia*. [Online]. Available: http://en.wikipedia.org/wiki/Entropy_(information_theory), [Accessed May 10, 2014].

[38] “Tails: The amnesic incognito live system.” [Online]. Available: https://tails.boum.org/index.en.html.

[39] R. Dingledine, N. Mathewson, and P. Syverson, “Tor: The second-generation onion router,” Naval Research Lab, Washington DC, Tech. Rep., 2004.

THIS PAGE INTENTIONALLY LEFT BLANK

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California